İSTANBUL TECHNICAL UNIVERSITY ★ INSTITUTE OF SCIENCE AND TECHNOLOGY

A RECOMMENDATION MODEL FOR SOCIAL RESOURCE SHARING SYSTEMS BASED ON TRIPARTITE GRAPH CLUSTERING

M.Sc. Thesis by
Yonca ÜSTÜNBAŞ

Department : Computer Engineering

Programme : Computer Engineering

Thesis Supervisor: Assoc. Prof. Dr. Şule GÜNDÜZ-ÖĞÜDÜCÜ

NOVEMBER 2011

İSTANBUL TECHNICAL UNIVERSITY ★ INSTITUTE OF SCIENCE AND TECHNOLOGY

A RECOMMENDATION MODEL FOR SOCIAL RESOURCE SHARING SYSTEMS BASED ON TRIPARTITE GRAPH CLUSTERING

M.Sc. Thesis by
Yonca ÜSTÜNBAŞ
(504081536)

Date of submission : **01 Nov 2011**
Date of defence examination: **03 Nov 2011**

Supervisor (Chairman) : **Assoc. Prof. Dr. Şule GÜNDÜZ-ÖĞÜDÜCÜ (ITU)**
Members of the Examining Committee : **Assoc. Prof. Dr. Ayşegül GENÇATA YAYIMLI (ITU)**
Assoc. Prof. Dr. Banu DİRİ (YTU)

NOVEMBER 2011

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

SOSYAL KAYNAK PAYLAŞIM SİTELERİ İÇİN ÜÇ PARÇALI ÇİZGE DEMETLEME YÖNTEMİ İLE ÖNERİ ÜRETME

YÜKSEK LİSANS TEZİ
Yonca ÜSTÜNBAŞ
(504081536)

Tezin Enstitüye Verildiği Tarih : 1 Kasım 2011
Tezin Savunulduğu Tarih : 3 Kasım 2011

Tez Danışmanı : Doç. Dr. Şule GÜNDÜZ-ÖĞÜDÜCÜ (İTÜ)
Diğer Jüri Üyeleri : Doç. Dr. Şule GÜNDÜZ-ÖĞÜDÜCÜ (İTÜ)
Doç. Dr. Ayşegül GENÇATA YAYIMLI(İTÜ)
Doç. Dr. Banu DİRİ (YTÜ)

KASIM 2011

FOREWORD

First, I would like to express my gratitude to my family for their continuous support. They have always supported me with love and patience.

I would like to express my deep appreciation and thanks for my advisor Assoc. Prof. Şule Gündüz Öğüdücü. I thank her for accepting me to her lab and for guidance on my thesis with patience and understanding. I also would like to thank Prof. Eşref Adalı for his guidance and encouragement.

Finally, I would like to thank to my friends and colleagues for their support and friendship.

September 2011

Yonca ÜSTÜNBAŞ
Computer Engineer

TABLE OF CONTENTS

Page

ABBREVIATIONS

TC **:** Tripartite Clustering
BC **:** Bipartite Clustering

LIST OF TABLES

Page

LIST OF FIGURES

Page

A RECOMMENDATION MODEL FOR SOCIAL RESOURCE SHARING SYSTEMS BASED ON TRIPARTITE GRAPH CLUSTERING

SUMMARY

One of the applications of Web 2.0 is social resource sharing systems. Some of these systems are using tagging approach, which is a kind of bottom-up classification technique when compared to hierarchies. Tagging is the term for assigning a personally chosen keyword to a piece of information. This approach has been popularized and has become an important feature of Web 2.0 Web. When hundreds of millions of users used this approach, these systems resulted with a knowledge representation called folksonomy. The term "folksonomy" describes a classification system derived from the practice and a method of collaboratively creating and managing tags to annotate and categorize content.

Recommendation engines help people to make decisions by providing options which the user may be interested in. Different recommendation systems use various methods, concepts and techniques from different research areas. The content and structure information extracted from folksonomies, makes it a good candidate for recommendation models.

When a social tagging system allows a user to bookmark a resource with a specific keyword, a tripartite relationship is built among user, resource and the keyword. As more people tag the same resources with different keywords, the resulting network of these tripartite relations, provides valuable information for generating recommendations. Tripartite clustering of this network, helps a recommendation system to understand the natural grouping in the dataset and enables it to identify the similar and different ones, so that, it can use this knowledge while training the model.

In this study, we implemented a web-based recommender system that uses a tripartite clustering algorithm for synchronous retrieval of cluster information of users, tags and resources. Resulting information of this algorithm enables our recommendation engine to make both Web page recommendations and tag recommendations.

The implemented model is experimented with a real dataset, which is gathered from one of the most popular social bookmarking systems, Delicious and results are evaluated by comparing to bipartite clustering. Simultaneous clustering of tripartite structured data provides more useful information for a recommender system, than bipartite clustering of the pairs, because of deciding the cluster contents with one more, different type of information source. Experiments show that tripartite clustering for Web recommendation outperforms bipartite clustering and it provides more information to let our engine making tag recommendations for the user.

SOSYAL KAYNAK PAYLAŞIM SİTELERİ İÇİN ÜÇ PARÇALI ÇİZGE DEMETLEME YÖNTEMİ İLE ÖNERİ ÜRETME

ÖZET

Web 2.0 ile gelen uygulama alanlarından biri de sosyal kaynak paylaşım sistemleridir. Bu sistemlerin bir kısmı hiyerarşilerle karşılaştırıldığında aşağıdan yukarı bir sınıflandırma tekniği olan etiketleme yaklaşımını kullanır. Etiketleme, bir bilgi parçasınına kişisel olarak seçilmiş bir anahtar kelimenin atanması işlemidir. Bu yaklaşım zaman içinde popülerleşti ve Web 2.0'ın en önemli özelliklerinden biri haline geldi. Yüz milyonlarca insanın bu yaklaşımı benimseyip kullanması sonucunda bu sistemler folksonomi adı verilen yeni bir bilgi temsil sistemi yaratmış oldular.

Öneri araçları, insanlara ilgilenme ihtimalleri yüksek olan seçenekler sağlayarak karar verme aşamasında fayda sağlar. Farklı araştırma alanlarında pek çok değişik konu, teknik ve method kullanan öneri modelleri geliştirilmiştir. Folksonomilerden çıkartılan bilginin içeriği ve yapısı, öneri modelleri için kaynak olarak kullanılmaya uygundur.

Bir sosyal etiketleme sistemindeki kullanıcının bir kaynağı bir anahtar kelime ile etiketlemesi ile kullanıcı, kaynak ve etiket arasında üçparçalı bir ilişki kurulmuş olur. Daha fazla insanın aynı kaynakları farklı kelimler ile etiketlemesi sonucunda oluşan üçparçalı ilişkilerin oluşturduğu ağ yapısı, öneri üretmekte kullanılabilecek değerli bilgiyi içerir. Bu ağın üç parçalı olarak demetlenmesi, bir öneri sisteminin, veri seti içindeki doğal gruplaşmaları anlamasını, farklı ve benzer olanları tanımlayabilmesini ve bu bilgi ile modelini eğitebilmesini sağlar.

Bu çalışmada, kullanıcı, kaynak ve etiketlerin demet bilgilerinin eşzamanlı edinimini sağlayan üç parçalı demetle algortiması gerçekledik. Bu algoritmayı kullanarak elde ettiğimiz bilgi, öneri modelimizin hem Web sitesi öneriminde hem de etiket öneriminde kullanılmasına olanak verdi.

Gerçeklenen sistemin en popüler etiketleme sistemlerinden biri olan Delicious'tan alınan veri seti üzerinde deneyleri yapıldı ve sonuçlar iki parçalı demetleme ile elde edilen sonuçlarla karşılaştırılarak değerlendirildi. Bir öneri modeli için üç parçalı verinin eşzamanlı olarak demetlenmesi, demet içeriklerine farklı bir bilgiyi daha katarak karar verdiği için, ikililerin iki parçalı demetlenmesi ile elde edilenden daha fazla faydalı bilgi sağladı. Deney sonuçları, Web sayfası öneriminde üç parçalı yapının iki parçalı yapıya göre daha iyi kesinlik sonuçlarına ulaşmasının yanında, öneri modelinin kullanıcı için tag önerimi de yapmasına olanak tanıyacak faydalı bilgiyi içerdiğini göstermiştir.

1. INTRODUCTION

Social resource sharing systems allow users to share their resources online, describe, and organize them by using tagging approach. Tags, other than working for the users as identifiers of the Web pages they assigned, have another assignment for social resource sharing systems as user specific interest identifiers. When a social bookmarking system takes a look at all tags of a user, it could have an opinion on the general interests of the user. These systems provide information about the users' basic interests, what they want to learn about, how they spend their time on the Internet, what they want to remember, what they find valuable, interesting, entertaining or educational. They can even give clue about the users' characters by identifying their interest on resources.

Other than giving general information on a user's basic interest, they define the interest of the user on a specific resource. Different users may have different interest on the same resource. For instance, a page about coding would be related with the keyword "work" for a programmer whereas an engineering student would define it as "educational". Someone else thinks that the page is about "geek stuff" and the other thinks that it is a good "source" for finding some materials. Besides, the same student would tag a Mathematics related Web site with the same tag, "educational" whereas a student who studies at Political Interactions would assign the tag "educational" to a Web page which has a completely different topic.

When a social resource sharing system allows users to tag resources with these personal keywords, a tripartite relationship is constructed among users, resources and tags. These systems result with a tripartite graph structured data collection called folksonomy which is a rich resource for data analysis, information retrieval, and knowledge discovery applications. Data mining is referred as a step in the knowledge discovery process consists of particular techniques for extracting models from data, and Web mining is the use of these data mining techniques to automatically discover and extract information from Web documents and services. Folksonomy mining is

the term of a branch of Web mining to discover useful knowledge and patterns from the folksonomy.

In this thesis, we implemented a recommendation model for social resource sharing systems using folksonomy mining techniques. A tripartite graph clustering technique is applied on the folksonomy dataset for analysing the connections and extracting the natural groupings in the dataset. The implemented recommendation model makes use of these resulting clusters to generate Web page and tag suggestions for the user.

This thesis is organized as follows: first in chapter 1, we provide a background about the folksonomy systems and recommender models and review the related work. In section 2, we enplane the tripartite structure of folksonomy and present the Tripartite Clustering Algorithm with its formulation and algorithm. Section 3 covers the data preparation and cleaning step of the knowledge discovery process. In Section 4, we presented the implemented recommendation model in detail. Section 5 includes experimental results and evaluation of these results. Finally, Section 6 concludes the work and provides directions for possible future work.

1.1 Folksonomy

Resource sharing systems were the first appearances of Web 2.0 applications. In time, some of these systems started to use tagging approach, which is a kind of bottom-up classification compared to hierarchies. Tagging is the term for assigning a personally chosen keyword to a piece of information. This approach then has popularized and has become an important feature of Web 2.0 applications. When hundreds of millions of users used this approach, these systems resulted with a knowledge representation called folksonomy. The word folksonomy stands for “taxonomy” created by “folks”, means user generated taxonomy to categorize web content. There is no hierarchy or a parent-child relationship between these users assigned terms. Unlike formal taxonomies they are not predetermined set of classification terms or labels, they are simply the set of terms that a group of users tagged content with [1]. Folksonomy can also be defined as a classification system that reflects the opinions of the public.

Well-known examples of folksonomy systems are Delicious, Flickr, CiteUlike, Youtube. Flicr and Youtube are photo and video management and sharing Web applications whereas Delicious and CiteULike are tools to organize Web pages. Delicious is described as a social bookmarks manager which allows people to save bookmarks online, share them with other people, and see what other people are bookmarking [2]. This online self-description also defines tags as "words people use to describe a bookmark" and explains how these tags allow users to describe and organize content with any vocabulary they choose. It is free to use the system after a registration. When the user wants to save a new bookmark, a form is presented to the user, which allows him to add tags and notes related to the bookmark. Tag field also includes a recommendation line below, which suggest tags to the user.

Some of the most popular tags (as of 15 August 2011) according to the system were "design, blog, video, software, tools, music, programming, webdesign, reference, tutorial, art, web, howto, photography, free, news, food, inspiration, linux". Some of these are technical subject tags, e.g., "linux, software" and some are genre or form descriptors, e.g. "photography, art". Some are more likely to reflect personal interest, e.g., "inspiration, free". Figure 1.1 represents the tripartite structure of folksonomy.

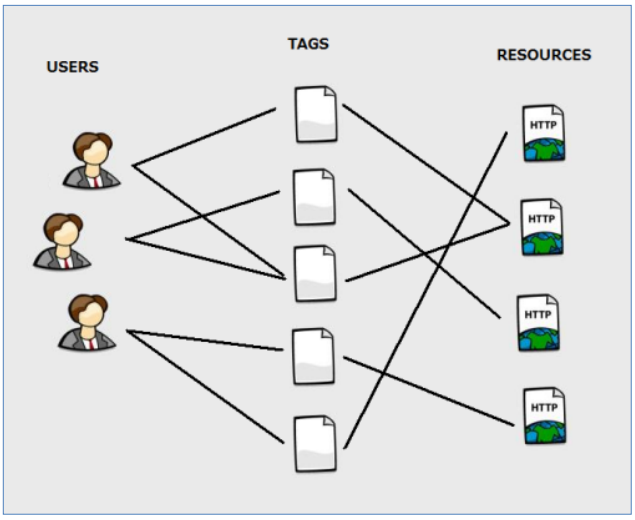

Figure 1.1: Tripartite structure of folksonomy.

Since there are no boundaries to classify an item, uncontrolled vocabulary in folksonomies leads to many limitations and weaknesses. When users apply the same tag in different ways, ambiguity can emerge. With the usage of synonyms, users can apply different tags for the same concept, e.g., “mac” and “apple”. Users may prefer assigning different inflections of words or may use different word forms, singular and plural, e.g., “mac”, “macintosh” and “car”, “cars”. Beside these kinds of ambiguous and inexact tags, overly personalized and misspelled tags make folksonomies chaotic and imprecise.

Flexibility, adaptability and serendipity are some of the important strenghts and advantages of the folksonomies. Their uncontrolled nature and organic growth can adopt very quickly to user vocabulary changes and needs without any significant cost. Providing social search and navigation results with the possibility of finding unexpected things in a general area and improves social serendipity by enabling social connections [3].

These characteristics of folksonomy make it a rich resource for data analysis, information retrieval, and knowledge discovery applications [4]. Folksonomy mining is the term of a branch of Web mining to discover useful knowledge and patterns from the folksonomy.

1.2 Recommender Systems

A recommender system is a system which generates recommendations to a group of users for items or products that might interest them. A major part of decision making involves the analysis of a finite set of alternatives in terms of some criteria. These systems simply help people to make decisions by providing alternatives. Most familiar examples for end users are the ones which suggest movies, books, music, videos and web pages. Today many popular Web applications and e-commerce sites include their own recommender systems. Amazon, NetFlix and MovieLens are well known examples of this area.

1.2.1 Categories of recommender systems

Recommender systems differ in the way they analyse data sources to measure similarity between users and items to identify matching pairs [5]. Collaborative Filtering systems analyse historical interactions alone whereas Content Based

Filtering systems are based on profile attributes. There are also hybrid systems that combine these two approaches.

Collaborative Filtering

Collaborative systems generate recommendations based on the past ratings of all users collectively. In other words they recommend items that are liked by other users with same interests. They collect user feedback, record of user preferences or extract them from user behaviours. Then they use these historical records to measure similarities and determine people with same interest. They are subdivided into two approaches; model-based and neighbourhood-based methods. Neighbourhood-based methods are also called memory-based since they work by choosing a subset of users based on their similarity to the active user. All users are assigned a weight and a prediction is computed from the most similar set of users. The other approach, model-based methods, provides recommendations by estimating parameters of statistical models for user ratings. These models use different machine learning algorithms such as clustering, Bayesian networks and rule-based approaches.

Content Based Filtering

Collaborative filtering recommenders threat all users and items as atomic units by only utilizing the rating matrix. More knowledge about a user can make a recommender system generate more personalized recommendations. For instance, demographic information of a user or genre of a movie might be more useful than just a rating value. Content Based Approaches provide recommendations by comparing representations of content describing an item to the one user interests. The difference between two approaches can be explained as, these systems select items based on the correlation between the content of the items as opposed to a collaborative system selects an item based on the correlation between people with similar preferences [6].

Hybrid Approaches

Hybrid recommendation systems combine many recommendation strategies and many input data sources in order to take advantage of both collaborative and content based filtering techniques and to provide better performance.

1.3 Web Page Recommender Systems

One type of recommender systems is called Web Page Recommender which is used for predicting the Web pages that are likely to be visited next to guide Web users in order to find information relevant to their needs [7]. These recommenders use Web mining approaches which use data mining techniques to gather and extract information from Web documents and services.

Web mining approaches are categorized into three areas according to the part of Web they are working on. These three categories are called Web Structure Mining, Web Usage Mining and Web Content Mining. Web usage mining is also called Web log mining since it deals with secondary data on Web. Web logs, proxy server logs, user queries, user profiles, cookies, registration and bookmark data are some of the examples of secondary data. Web content mining works on analysing the text and multimedia content on the Web whereas Web structure mining usually works on link structures between Web pages. Web recommender models generally combine techniques from all categories in order to provide better performance.

A Web page recommender systems basic goal is to determine the user's interest early while he/she is browsing the Web site. This information is then used to keep the user browsing the Web site for a longer time or increasing the possibility of visiting the page again by helping his/her to access what he/she is looking for. This would make the Web site more competitive by increasing its performance. Besides advising users about the Web pages they might be interested in, Web recommender systems can also improve the Web performance through caching operations. Caching and pre-fetching the right Web pages would help to increase the speed of the Web site significantly which would also might be useful to keep the users on the Web site. Besides, recommending well matched advertisements for users is important for advertising agencies which pay these Web sites.

1.3.1 Recommendation process

The recommendation process of a recommender system is parallel to the well known data mining process. It consists of three stages; first one is data collection and pre-processing, second is pattern discovery and analysis on this data, and the third stage is the recommendation stage. As mentioned earlier, Web mining techniques works on many different types of data, gathered from Web documents and services. In general,

recommender systems make use of Web server logs that hold the browsing history records of the users. Data collection and pre-processing stage covers the gathering and preparation of these server logs. In the pattern discovery and analysis stage, data mining methods are employed on the clean and structured data. The most common techniques employed in this stage are Markov models, association rules generating, sequential pattern generation, Collaborative Filtering and clustering user sessions. These methods make the system to analyse usage patterns and generate recommendations automatically. Recommendation stage is the stage when the tracks of active user sessions are kept and the recommendation model generates recommendations using the requests of the active user. This step is online in order to get requests of active user and provide recommendations before the Web page is sent to the client browser.

Data Collection and Pre-processing

Recommender systems are categorized by how they model users: explicitly or implicitly [7]. Explicit user modelling is an approach in which, users submit their own personal information and feedback about Web pages in the form of ratings whereas in implicit user modelling, the user model is built by observation and data mining methods. The second approach is a more realistic way since users doesn't generally spend time for rating pages. The goal of this approach is to minimize user collaboration and still be effective by extracting valuable information about the user interest from the collected data.

The implicit user models are formed using Web log data, the content of web pages, the structure of the Web site and user queries. These data types can be categorized as content data, user data, usage data, and structure data. Content data covers all objects and relations that are visible to users. The content organization within the Web site corresponds to the structure data. User data is collected via user interaction which represents a user profile for a recommendation system and the usage data refers to the Web server access logs that are recorded by Web servers.

The pattern discovery and analysis stage of recommendation process is the step in which different kinds of data mining techniques are employed. These algorithms must work on structured, reliable, integrated dataset in order to produce effective results. For generating this kind of dataset, various pre-processing approaches are applied to the obtained data. User and session identification and page time

calculation are examples of data pre-processing methods that are applied on Web usage data. Various types of noise elimination methods are also applied on Web content and structure data in the pre-processing step. The pre-processing methods applied for this research are explained in detailed in the following sections.

Pattern Discovery

Recommender systems employ different pattern discovery methods for two purposes: user modelling and recommendation. These methods are generally combinations of different data mining techniques, which can be categorized as: collaborative filtering, clustering, association rules, sequential patterns and semantic Web.

Collaborative filtering systems work on user ratings that are collected from visitors. These systems construct a user×item rating matrix and use this matrix to predict users interest on an item, or in this case on a Web page. Collaborative filtering systems are less efficient on predicting the next Web page the user will visit since they ignore the sequence of page requests while modelling the behaviour of a user.

Association rule mining techniques extract set of Web pages that are accessed together with a support value exceeding a specified threshold. Association rules are generated from the user sessions to help a recommendation model find out the frequently accessed Web pages. Resulting pages are then used by the engine to make recommendations to the user whose current active session matches them. Association rule mining techniques do not consider the sequence of visiting Web pages since frequent page sets are not ordered. As a result, they may not be effective on predicting the next request of the user.

Page requests of user sessions are recorded in Web server logs. Sequential pattern discovery methods analyse this information to capture the Web pages that are frequently visited by users in the order that they were visited. Sequential pattern mining is also employed on many areas other than extraction of Web access patterns such as analysis of DNA sequences, natural disasters, customer purchase behaviours and etc. Markov-based models are widely used for modelling sequential processes such as browsing a Web site. These models are more capable of predicting the next request of a Web user since they consider the sequence of requests [7].

The goal of semantic Web techniques is to combine domain knowledge with the Web mining process by mapping the content of a Web site into ontology. Since creating

ontology from Web sites is a difficult task, it is harder to generate recommendations at the Web page level when compared to the other pattern discovery techniques. However this approach has some other advantages that make it preferable. Dynamically generated Web sites do not generally contain enough navigational patterns to work on for analysis since individual pages are not frequently requested. Semantic Web mining techniques could handle the problem of recommendation for these Web sites by providing domain knowledge. Also, this approach may solve the cold start problem for new pages. Still, these techniques demand high level of proficiency of expert and are domain dependent due to the use of domain ontology.

Clustering is a technique to extract natural groupings with similar characteristics in a dataset. Web mining techniques work on three kinds of clusters on Web usage data: user clusters, session clusters and page clusters. When there is not enough information about users, sessions and pages are clustered. Session clusters include sessions in which users have similar access patterns. The information gathered from clustering is used by a recommendation engine by depending on the idea that users with similar characteristics are more likely to have similar interest on a Web page. In most of the Web page prediction systems clustering is used with other data mining techniques.

There are also different approaches for evaluating the results of a Web page recommendation system. For instance, the calculation of accuracy of a Web page recommendation model differs according to the generation of recommendation set. Besides, the accuracy can be evaluated in terms of precision and coverage metrics [7]. Recommendation diversity and popularity metrics are also evaluated when accuracy is not a complete indicator of the recommendation model [7]. Appropriate evaluation metrics are chosen depending on several parameters, such as the users utilities on the model, the comparability of the previous studies, whether the model considers the order of the page requests or not and so on.

Different types of pre-processing and pattern extraction methods and some evaluation metrics for Web page recommendation are presented above. Several pre-processing methods and a clustering technique are applied in this research for prediction of Web pages and tags.

1.4 Tag Recommender Systems

Other than recommending resources, recommender systems can also suggest tags for assigning a label to a resource. Recommending tags to users improves the usage of social tagging systems by increasing the number of tagged resources. Also, tag suggestions can help to generate a common vocabulary among users.

Tag recommender systems can be categorized into three classes: collaborative, content based and graph based tag recommender systems [8]. Content based tag recommender systems make use of the information gathered from resources and users, such as content of resources and demographic information of a user for suggesting tags. Collaborative tag recommender systems analyze metadata, provided by users to resources, indicating the relevance of tags to a specific resource. Hybrid approaches are also exist but not categorized into a different class for tag recommendation. The third approach, graph based system, represents folksonomy as a graph and generate tag recommendation by using this model.

These systems are also divided into two classes by using different criteria, the relevance of tags to be suggested: personalized and not personalized tag recommender systems [9]. Not personalized systems do not consider users' tagging habits and generate same suggestions for different users who are interested in the same resource. In contrast, personalized systems do not ignore users' personal way to classify resources and suggest the most relevant tag for different users.

1.5 Related Work

One of the first analyses on folksonomy mining relies on creating frequent sets and uses apriori algorithm to extract the association rules from folksonomy data [10]. Right after the analysis, many recommendation systems are implemented based on these social bookmarking information [11][12]. The first recommender systems which are based on folksonomies, simply used tags as topics, that, users are interested in.

Tripartite graph structure of folksonomies leads to research on formalization of this context [13][14][15]. Most of the research on this topic, mainly studies on popular folksonomy systems. Comprehensive analysis presented the structure and properties of the most popular examples of these systems, namely Bibsonomy and Delicious

[16][4]. They observed and noted that the tripartite hyper graphs of folksonomies of these systems are highly connected and that the relative path lengths are low. Many of the studies have dealt with the major problems of folksonomy mining, such as tag redundancy and ambiguity [11][12]. Some of them focused on identifying solutions to these problems [17]. Due to the fact that Web documents which correspond to the same meaning of a tag tend to be grouped together to form clusters, hierarchical clustering of tags based on their affinity levels, were presented as a solution. Another suggested solution was extracting bipartite clustering of users and documents to help to disambiguate tags [13].

In addition to Web page recommendation, the information extracted from folksonomy data can also be used for tag recommendation. A survey titles recommender algorithms as tagommenders that predict users' preferences for items based on their preferences for tags [18]. An early study on tag recommendation, explains that recommending tags can serve various purposes, such as: increasing the chances of getting a resource visited, reminding a user what a resources topic is and extracting the vocabulary across the user [19]. A recent study on tag recommendation, proposed an improved version of k-means for social tagging data [20]. They first used social annotation data to expand the keyword vector space model of k-means clustering and then applied the links involved in social tagging network to enhance the clustering performance.

Another study proposed a highly-automated novel framework for real-time tag recommendation by representing the triplets as two bipartite graphs and defined some evaluation metrics to measure the effectiveness of their algorithm [21].

A recent study adopted a model for analyzing the structure of k-partite graphs for clustering tripartite network of a real world social tagging dataset [22].

The studies explained above either gives contextual information or propose different models and algorithms to work on tripartite structure of folksonomies. Recommender systems in this context, that are developed so far, have rarely taken the advantage of tripartite structure of folksonomies simultaneously In this study, we implement a clustering method called "tripartite clustering" [23] which cluster the three types of nodes (resources, users and tags) simultaneously based on the links in the social tagging network. This method is compared with content based k-means approach and it has been proven that it significantly outperforms whereas producing much more

useful information. It is a novel method and its advantage of synchronized retrieval of clusters also enables us to develop a recommender system for both Web pages and tags.

1.6 Goal of the Thesis

Social resource sharing systems allow users to upload their resources and label them with a keyword from their own vocabulary. The users not only organize and share their resources but also explore new resources by visiting their friends or other people's uploads in this kind of a social environment. The goal of this work is to build a recommendation model for social resource sharing systems by making use of the tag information. The model can be used to suggest Web pages that the user might be interested in as well as for generating tag suggestions when he/she decides to tag a Web page. It can be presented as a Web page and tag recommender model for resource sharing systems. The intend of this work is to help a user in these systems by providing options while tagging and exploration of new resources. By analysing the connections between the users, resources and tags simultaneously, the model recommends resources and keywords that might be useful or interesting for the user. The model helps them to find out new resources about a specific information they were searching for or new area of interests that they would like. Besides its help on the users of the resource sharing system, the model also helps the system itself. It improves the performance and preferability of the system by satisfying the user needs and it increases the probability of having tagged resources by proving tag recommendations. Tag recommendation part of the model can also be used for creating a common vocabulary among users.

2. TRIPARTITE CLUSTERING

This chapter explains the tripartite network structure of folksonomies in detail and introduces the implemented method in this thesis. Tripartite clustering model is presented with its formulation and algorithm.

2.1 Tripartite Structure of Folksonomies

When a social tagging system allows a user to bookmark a resource with a specific keyword, a tripartite relationship is built among user, resource and the keyword. This keyword, called a tag, represents the meaning of that page to the user. The information, that tags sustain, creates a significant interconnection between user and resource. A folksonomy, the data in a social tagging system, is defined as follows [22] .

A folksonomy F is a tuple F = (U,T,D,A), where U is a set of users, T is a set of tags, D is a set of Web documents, and $A \subseteq U \times T \times D$ is a set of annotations.

Since a folksonomy has three elements, which all have interconnections between them, three different types of bipartite graphs can be extracted from it. For example when we take the documents and the tags, they are connected by users associated them. Or when we only take the users and tagged documents they are connected by tagging operation. If it is represented as a weighted graph, its weights can be the tag count given to the document from that user. For the first example, the weights of the bipartite graph of documents and tags can be represented with the number of users, who associated them.

The term ‘resource’ is a common term for defining tagged element in a tripartite network. Some of the examples of social tagging systems are Flickr, CiteULike or Del.ici.ous. In Flickr, resource is a photograph whereas in our case, in folksonomies, a resource means a Web document. Again, in this context, because we experiment a social tagging system, it is expected that a Web resource is tagged with a distinct tag only one time. Because it is meaningless for a user to tag the same resource with the

same keyword for two times, the tripartite network of folksonomies is not a weighted graph, unlike its bipartite portions.

2.1.1 Bipartite graphs in folksonomies

As mentioned, interconnections between three elements of a social tagging system, includes three different types of bipartite graphs. These graphs are formulated and explained similarly in [13].

The bipartite graph of tags and resources can be denoted as follows. If a user has assigned a tag to a resource, an edge exists between the tag and the resource. Links in the graph of tags and resources are weighted by the number of users that have connected them.

$$TR_u = \langle T \times \mathrm{R} \big| E^{(TR)} \rangle, E^{(UR)} = \{(t,r) | (u,t,r) \in A\} \quad \textbf{(2.2)}$$

The graph of users and resources can be denoted similarly as follows. If a user has assigned a tag to a resource, an edge exits between the user and the resource.

$$UR_t = \langle U \times R \big| E^{(UR)} \rangle, E^{(UR)} = \{(u,r) | (u,t,r) \in A\} \quad \textbf{(2.2)}$$

The graph of users and tags are denoted in a similar way. If a user has assigned a tag to a resource, an edge exists between the user and the tag. Links in the graph of users and tags are weighted by the number of resources that the user assigned that tag.

$$UT_d = \langle U \times T \big| E^{(UT)} \rangle, E^{(UT)} = \{(u,t) | (u,t,r) \in A\} \quad \textbf{(2.3)}$$

2.2 Tripartite Clustering Model

In this thesis, we consider the problem of recommending resources or tags to users based on tripartite structure of folksonomies. We implement the method proposed in [23] for tripartite graph clustering, which can cluster the three types of data objects simultaneously and produces much more useful information for a recommendation system. Because similar users are more likely to perform similar behaviours, the resulting clusters of this method can be used for both Web page recommendation and tag recommendation.

Model Formulation

Similar to folksonomy definition, a social tagging system can be denoted as follows for formulating the tripartite graph clustering algorithm [23]:

$$TN = \left(U, R, T, E^{(UR)}, E^{(UT)}, E^{(RT)}\right) \quad \textbf{(2.4)}$$

where U is a set of users, R is a set of resources and T is a set of tags. $E^{(UR)}$ is the undirected links between users and resources. $E^{(UT)}$ is the undirected links between users and tags. $E^{(RT)}$ is the undirected links between resources and tags.

Each node in the network can be denoted by its link vector to the other type of nodes. For example, a resource can be represented by two link vectors, first one includes the weight vector of users, and the second one includes weight vector of tags. These vectors are formulated in the study as follows:

$$R_i^{(U)} = \langle y_{ih}^U | h = 1,2, \dots, |U| \rangle \quad \textbf{(2.5)}$$

r_i's user link vector, where y_{ih}^U is the weight assigned to the link from r_i to u_h and |U| is the size of set of users.

$$R_i^{(T)} = \langle y_{ij}^T | j = 1,2, \dots, |T| \rangle \quad \textbf{(2.6)}$$

r_i's tag link vector, where y_{ij}^T is the weight assigned to the link from r_i to t_j and |T| is the size of set of tags.

The weights are equal to one because it is expected that a user assigns the same tag to the same resource only one time.

This kind of formulation enables the separate clustering of three types of nodes. Vector Space Model based clustering algorithms can cluster three types of nodes with these vectors. But this would ignore the interconnections between the different types of nodes. For synchronous clustering of all types of nodes, Tripartite Clustering Algorithm introduces an iterative approach like k-means.

This algorithm has a similar approach with k-means. Following the random initialization of cluster numbers of nodes, nodes are iteratively replaced into the clusters depending on their distance to the centres of clusters. Different from k-

means, the distance between the node to be replaced and the centre of the cluster, is calculated by cosine similarity of the two types of node vectors. The other difference of this approach is the calculation of centroids of the clusters. To take into account the interactions among the cluster structures of different types of nodes, a centroid of a cluster is calculated by not only considering the nodes of the current cluster but also by taking into account the other two types of nodes. Figure 2.1 explains how the centroid of a resource cluster is affected by the cluster structures of the user nodes.

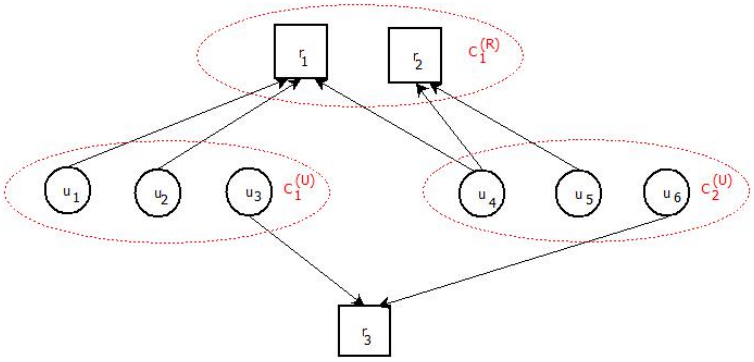

Figure 2.1: The distance between a resource node and the centroid of a resource cluster is affected by cluster structures of the user nodes.

The formula presented in Eq. 2.7, explains the calculation of the centroid of a resource cluster based on user link vectors at dimensionu_{μ}.

$$centroid_{m\mu}^{U} = \frac{\sum_{r_i \in C_m^{(R)}, u_{hi} \in C_l^{(U)}} y_{ih}^{(U)}}{|C_m^{(R)}| * |C_l^{(U)}|}, (u_{\mu} \in C_l^{(U)}) \quad \textbf{(2.7)}$$

,where $C_l^{(U)}$ $(1 < l < k_U)$ is the cluster that user U_{μ} belongs to. u_h represents a user node in cluster $C_l^{(U)}$, and r_i is a resource node in $C_m^{(R)}$. $y_{ih}^{(U)}$ is the weight of the link from r_i to u_h.

The centroid of the resource cluster is calculated similarly based on tag link vector. The distance from resource r_i to the centroid of a resource cluster $C_m^{(R)}$ can be calculated as follows:

$$d(r_i, centroid_m) = a * d\left(R_i^{(U)}, centroid_m^{(U)}\right) + (1-\alpha) * d\left(R_i^{(T)}, centroid_m^{(T)}\right) \quad \textbf{(2.8)}$$

,where $d\left(R_i^{(U)}, centroid_m^{(U)}\right)$ denotes the distance between r_i and the centroid of $C_m^{(R)}$ based on the resources' user link vectors; and $d\left(R_i^{(T)}, centroid_m^{(T)}\right)$ represents the distance based on the resources' tag link vectors. α quantifies the influence of the resource's user link vector on its clustering. Distance of other type of nodes to the centroids of clusters can be calculated similarly. In this study, we used cosine similarity to calculate the distance. The algorithm for tripartite graph clustering model is presented in Figure 2.2.

Input:

The social tagging network *TN* = (*U*,*R*,*T*,*E*(*UR*) ,*E*(*UT*) ,*E*(*RT*)); The cluster numbers of resource nodes, user nodes, and tag nodes.

Output:

The cluster assignment of resources, users and tags

Method:

Initialize cluster assignment: assign each node to a random cluster;

Repeat:

For each type of the nodes **do**

Calculate the centroid of each cluster based on the link features of its cluster members and the cluster structures of other two types of nodes from last iteration, as defined in equation (2.7);

For each resource (user or tag) node **do**

Calculate the distance between the node and the centroid of each resource (user or tag) cluster according to equation (2.8);

Reassign the current node to the closest cluster based on the distance value.

End For

End for

Until (The assignments no longer change **OR** Iteration Number

>=Threshold)

Figure 2.2: Algorithm of Tripartite Clustering Model.

3. DATA PREPERATION AND CLEANING

In this work, we use a real dataset obtained from Delicious website [24]. Delicious is an online social bookmark sharing system which allows people to save and share bookmarks with other people. It is basically a tool to organize Web pages. Users are allowed to describe their bookmarks with their own vocabulary.

Delicious provides data feeds for news readers and third party applications. [25] The Web site has an RSS feed of site-wide bookmarking activity. RSS is a Web feed format that is used to publish frequently updated Web content. Crawling tools are used to collect datasets from these feeds. Some of these previously crawled datasets are available on the Web for academic use. [26]

In this work, we use a previously crawled dataset which is in JSON format. [24] JSON, or JavaScript Object Notation is a lightweight, text based, human readable data interchange format. Figure 3.1 presents an example of a single line in the dataset. Each line of the dataset represents a single entry at a specific time. Each entry in the system is a bookmarking process of users for a single Web site. The presented bookmarking entry includes its entrance date and time, its author, link of the Webpage, title of the Webpage, tags assigned to it and so on.

For this thesis, we extracted the useful key-value pairs from this JSON formatted lines by eliminating the unnecessary fields. The implemented system for the thesis works on links between users, resources and tags on a social resource sharing system. So we extracted the author, link and tag values in an entry. Each entry is exported to the new file with a line for each of its tags and they are ordered by date. Each line in the file of the new dataset is in the following format, [user, url, tag]. Table 3.1 presents a sample of the extracted dataset.

Table 3.1: Sample dataset extracted from the original JSON formatted data.

User	Resource	Tag
phlaff	http://www.blurb.com/	Livre-photo
phlaff	http://www.blurb.com/	livre
phlaff	http://www.blurb.com/	photo
phlaff	http://www.blurb.com/	book-photo
phlaff	http://www.blurb.com/	book
eto	http://www.hanyuedu.net/	??????
ibbertelsen	http://brainz.org/15-coolest-cases-biomimicry	Biomimicry
ibbertelsen	http://brainz.org/15-coolest-cases-biomimicry	technology
ibbertelsen	http://brainz.org/15-coolest-cases-biomimicry	design
ibbertelsen	http://brainz.org/15-coolest-cases-biomimicry	biology
ibbertelsen	http://brainz.org/15-coolest-cases-biomimicry	engineering
ibbertelsen	http://brainz.org/15-coolest-cases-biomimicry	tecintense
ibbertelsen	http://brainz.org/15-coolest-cases-biomimicry	meft3102
crpurgas	http://www.allforgood.org/	volunteer
crpurgas	http://www.allforgood.org/	nonprofit
crpurgas	http://www.allforgood.org/	web2.0
crpurgas	http://www.allforgood.org/	com498
crpurgas	http://www.allforgood.org/	mie310
tsypa	http://www.businessdelo.ru/	startup
neosolo	http://www.massage-zen-therapie.com/radio-zen/	massage
neosolo	http://www.massage-zen-therapie.com/radio-zen/	music
neosolo	http://www.massage-zen-therapie.com/radio-zen/	zen
tortugax	http://schneestern.livejournal.com/157027.html	gerard/mikey/pete
tortugax	http://schneestern.livejournal.com/157027.html	gerard/mikey
tortugax	http://schneestern.livejournal.com/157027.html	fob
tortugax	http://schneestern.livejournal.com/157027.html	mcr
tortugax	http://schneestern.livejournal.com/157027.html	bandom
tortugax	http://schneestern.livejournal.com/157027.html	fic
tortugax	http://schneestern.livejournal.com/157027.html	nc-17
rocha	http://puredata.info/	processing
rocha	http://puredata.info/	audio
rocha	http://puredata.info/	puredata
rocha	http://puredata.info/	software
rocha	http://puredata.info/	video
rocha	http://puredata.info/	sound
bellebelle2	http://delanach.livejournal.com/10606.html	sam/dean
bellebelle2	http://delanach.livejournal.com/10606.html	nc-17
bellebelle2	http://delanach.livejournal.com/10606.html	pre-series
bellebelle2	http://delanach.livejournal.com/10606.html	supernatural
georgeandre	http://www.thefedoralounge.com/	fashion
georgeandre	http://www.thefedoralounge.com/	vintage
georgeandre	http://www.thefedoralounge.com/	retro
georgeandre	http://www.thefedoralounge.com/	clothing
georgeandre	http://www.thefedoralounge.com/	forum
georgeandre	http://www.thefedoralounge.com/	style
georgeandre	http://www.thefedoralounge.com/	clothes
georgeandre	http://www.thefedoralounge.com/	men

```
{
        "updated": "Sun, 06 Sep 2009 05:20:34 +0000",
        "links":
                [{
                        "href": "https://club.nintendo.com/",
                        "type": "text/html",
                        "rel": "alternate"
                }],
        "title": "Club Nintendo | Nintendo Member Rewards",
        "author": "adsilveiras",
        "comments": "http://delicious.com/url/9670a0a7bfe1ece8a19ae77112ff96c0",
        "guidislink": false,
        "title_detail":
                {
                        "base": "http://feeds.delicious.com/v2/rss/recent?min=1&count=100",
                        "type": "text/plain",
                        "language": null,
                        "value": "Club Nintendo | Nintendo Member Rewards"
                },
        "link": "https://club.nintendo.com/",
        "source": {},
        "wfw_commentrss": "http://feeds.delicious.com/v2/rss/url/9670a0a7bfe1ece8a19ae77112ff96c0",
        "id": "http://delicious.com/url/9670a0a7bfe1ece8a19ae77112ff96c0#adsilveiras",
        "tags":
                [
                        {
                                "term": "Games",
                                "scheme": "http://delicious.com/adsilveiras/",
                                "label": null
                        },
                        {
                                "term": "Nintendo",
                                "scheme": "http://delicious.com/adsilveiras/",
                                "label": null
                        },
                        {
                                "term": "Wii",
                                "scheme": "http://delicious.com/adsilveiras/",
                                "label": null
                        }
                ]
}
```


Figure 3.1: Example JSON representation of a bookmark on Delicious.

3.1 Data Cleaning

The dataset is then imported to a Database Management System for other pre-processing and cleaning steps. An id is assigned to each unique user, url and tag for increasing the speed of operations. Several filtering methods are applied in order to remove improper values. Entries that include empty fields, fields which completely composed of punctuation marks, signs or numbers are filtered. Sql queries are used for filtering operations.

In folksonomy, there are no boundaries to assign a tag to an item. This provides adaptability, flexibility and in the meanwhile, causes lack of control and leads to

some problems in folksonomy mining. Uncontrolled vocabulary of tags may lead to inappropriate interconnections between items when they include homonyms (the same tags used with different meanings) and synonyms (multiple tags for the same concept). Also users may prefer to assign different inflections of words as tags. Beside these kinds of ambiguous and inexact tags, over personalized and misspelled tags make folksonomies chaotic and imprecise. These characteristics of folksonomy lead to negative effects on the analysis. In order to reduce these effects and to get a more connected and semantically related sample, Porter stemming algorithm is applied on tags in the data pre-processing and cleaning step. It is a term normalization process for removing the commoner morphological and inflexional endings from the words in English [27]. Table 3.2 presents a sample of the main dataset after pre-processing steps and Table 3.3 presents some of the tags before and after the stemming process.

As proposed in other studies [20] the systems with graph partitioning work efficient on a non-sparse, highly connected, tripartite structured dataset. We filter the dataset in order to get a sample with these properties in the pre-processing step. Frequent users, resources and tags which occur at least ten times are extracted from the original dataset. This filtering process is applied after stemming operation since the count of unique tags decreases after stemming. For instance; the words photo, photos, photograph, photographs, photography are presented with the word “photo”. This also changes the statistics about the links between user, resources and tags such as; the number of times a link is assigned or the number times a resource is annotated with the same tag.

Since the implemented system for this thesis works on the links between users, resources and links, the model makes use of some tables, which include frequency information, other than the main dataset formatted in the [user, resource, tag] form. These tables are generated from the main dataset and they mostly provide statistical information for the model such as the tagging count of each resource, the number of times a tag is assigned to. These datasets are generated for increasing the speed of the model by eliminating some of the calculations. By storing statistics, online workload of the recommendation model is transferred to the offline process of the model. These datasets do not need any pre-processing operations because they are generated from the main dataset and they are generally composed of ids and counts.

Table 3.2: Sample dataset after pre-processing steps.

User	Resource	Tag
jtorresonline	http://www.smugmug.com/	photographi
jtorresonline	http://www.smugmug.com/	blog
jtorresonline	http://www.smugmug.com/	design
darkness	http://commons.apache.org/sandbox/csv/	java
darkness	http://commons.apache.org/sandbox/csv/	apach
darkness	http://commons.apache.org/sandbox/csv/	csv
darkness	http://commons.apache.org/sandbox/csv/	api
darkness	http://commons.apache.org/sandbox/csv/	librari
Jive	http://www.revostock.com/	stock
Jive	http://www.revostock.com/	project
Jive	http://www.revostock.com/	aftereffect
Jive	http://www.revostock.com/	vfx
Jive	http://www.revostock.com/	video
Jive	http://www.revostock.com/	audio
Jive	http://www.revostock.com/	fx
Jive	http://www.revostock.com/	resourc
Jive	http://www.revostock.com/	produc
Jive	http://www.revostock.com/	post
Jive	http://www.revostock.com/	busi
Jive	http://www.revostock.com/	sell
Jive	http://www.revostock.com/	motion
Jive	http://www.revostock.com/	librari
Jive	http://www.revostock.com/	databas
Jive	http://www.revostock.com/	shop
Jive	http://www.revostock.com/	servic
Jive	http://www.revostock.com/	webbas
Jive	http://www.revostock.com/	download
grueda	http://ebookmall.com/	book
grueda	http://ebookmall.com/	shop
grueda	http://ebookmall.com/	ebook
francophilenz	http://www.language-archives.org/	languag
francophilenz	http://www.language-archives.org/	archiv
kantel	http://www.ibegin.com/labs/wp-lifestream/	wordpress
kantel	http://www.ibegin.com/labs/wp-lifestream/	flickr
kantel	http://www.ibegin.com/labs/wp-lifestream/	aggreg
kantel	http://www.ibegin.com/labs/wp-lifestream/	plugin
kantel	http://www.ibegin.com/labs/wp-lifestream/	wordpressplugin
kantel	http://www.ibegin.com/labs/wp-lifestream/	socialnetwork
kantel	http://www.ibegin.com/labs/wp-lifestream/	lifestream
kantel	http://www.ibegin.com/labs/wp-lifestream/	plugin
kantel	http://www.ibegin.com/labs/wp-lifestream/	rss
KeithVallis	http://dotsub.com/	video
KeithVallis	http://dotsub.com/	languag
KeithVallis	http://dotsub.com/	movi
KeithVallis	http://dotsub.com/	movi

We divided the dataset into training and test sets using sessions in order to generate recommendations for user profiles. A detailed user session analysis is not conducted. A user session is the presence of a user with a specific IP address in a period of time. What we did for a user session is to prepare it as a user bookmarks a fixed number of pages in a specific time period.

Original dataset contains of 213,628 entries that are created by users. After applying pre-processing steps, the resulting dataset is divided into training and test sets with % 70/% 30 ratios. In order to be able to evaluate the performance of the recommender model, dataset is divided into test and training sets by implying the following technique. The entries of each user are divided into two divisions. The first 70% of the entries of each user u is separated as the training set and the remaining part of her entries is separated as the test set. At the end of this step, the training set includes 12,998 rows, representing transitions between 1054 unique users, 900 unique resources and 668 unique tags. Each row corresponds to a tagging operation of a user in the following format, [user url tag] Test set, in the same format, contains 5,571 rows, representing transitions between 552 unique users, 629 unique resources and 579 unique tags.

Table 3.3: Some of the tags before and after the stemming process.

Unstemmed Tag	Stemmed Tag
h800_block3_2009	h800block32009
howto	howto
tips	tip
media	media
iphone	iphon
apps	app
scheduling	schedul
tools	tool
calendar	calendar
events	event
web2.0	web20
planning	plan
productivity	product
collaboration	collabor
REFERENCE	refer
congress	congress
politics	polit
liberals	liber
democrats	democrat
twitter	twitter
espiritual	espiritu
crystal	crystal
iphone	iphon
facebook	facebook
api	api
webdev	webdev
CSS	css
reference	refer
humor	humor
references	refer
comedy	comedi
Magazines	magazin
php	php
symfony	symfoni
rss	rss
browser	browser
web	web
curl	curl
feed	feed
web2.0	web20
redes	rede
regexp	regexp
music	music
lists	list

4. RECOMMENDATION MODEL BASED ON TRIPARTITE CLUSTERING APPROACH

The process of a Web page recommendation system was explained briefly in Chapter 1. The process of the recommendation system implemented in this thesis differs from the explained system by brunching into two recommendation systems in the last step; Web page recommendation and tag recommendation. First stage of the recommendation system, data preparation and cleaning, is explained in detail in Chapter 2. Different types of pre-processing and pattern extraction methods and some evaluation metrics for Web page and tag recommendation are presented in the Introduction section. Chapter 3 covered the implemented method for the second stage of the recommendation process. Tripartite clustering method is applied for extracting the natural groupings of users, resources and tags simultaneously in the pattern discovery and analysis step. For the recommendation step, resulting clusters of the model are used for generating well-matching suggestions to the users.

4.1 Proposed Model

A recommendation model based on clustering simply recommends Web pages and tags relying on the idea that similar users with similar interest are more likely to have similar navigational patterns. Tags, other than working for the users as identifiers of the Web page they are assigned, have another assignment for social resource sharing systems as user specific interest identifiers. When a social bookmarking system takes a look at all tags of a user, it could have an opinion on the general interests of the user. These systems provide information about the users' basic interests, what they want to learn about, how they spend their time on the Internet, what they want to remember or what they find valuable, interesting, entertaining or educational. They can even give some clue about the users' characters by identifying their interest on resources.

Other than giving general information on a user's basic interest, they define the interest of the user on a specific resource. Different users may have different interests

on the same resource. For instance, a page about coding would be related with the keyword “work” for a programmer whereas an engineering student would define it as “educational”. Someone else thinks that the page is about “geek stuff” and the other thinks that it is a good “source” for finding some materials. Besides, the same student would tag a Mathematics related Web site with the same tag, “educational” whereas a student who studies at Political Interactions would assign the tag “educational” to a Web page with a completely different topic. The recommendation model implemented for this thesis analyses all of these connections to generate personalized recommendations for these users.

Tripartite clustering model enables the system to extract user profiles not only by grouping the Web pages they visit but also by their personal interests on the Web pages. Simultaneous grouping, results in clusters of users, resources and tags. The recommendation model makes use of the clusters of users as user profiles. Each user is presented to the system by its matching cluster. Once a user is recognized by the system, then the Web pages, he/she might like to visit, and the tags, he/she would prefer to assign to those pages, can be predicted by the system. Resulting clusters of Web pages are used for generating Web page recommendations by creating candidate Web page sets from each cluster. Most frequently visited Web pages of each cluster are extracted as a candidate set for the users who are placed in the same cluster. Candidate sets contain the Web pages which are defined as “most likely to be visited” by the users in the same cluster. Candidate sets of tags are also created from resulting tag clusters and they are suggested when the user decides to tag a Web page. Figure 4.1 visualizes methodology of the implemented recommender system.

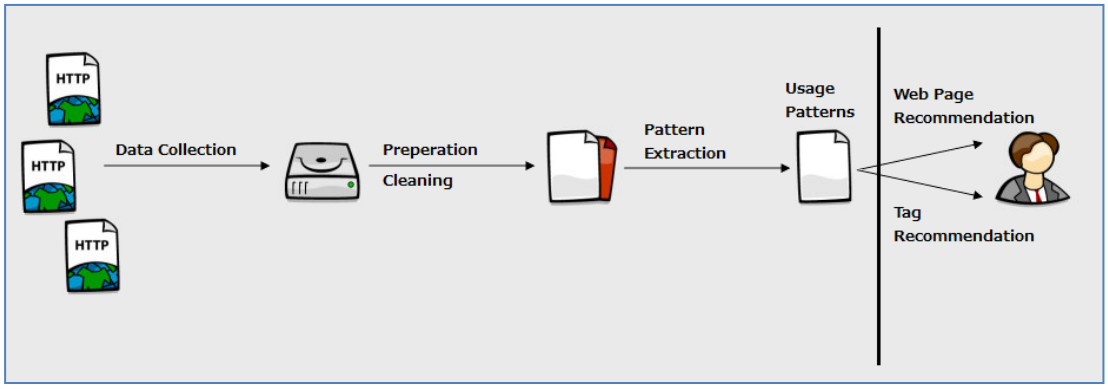

Figure 4.1: Methodology of the recommender system.

The recommendation model for Web pages and tags differ in some ways. For instance; in tag recommendation, recommender model looks for the resource profile if it can't find the user in the system. They also use different metrics for evaluating results. Recommendation models are explained separately in the following section.

4.2 Web Page Recommendation Model

Synchronized clustering of folksonomy data results with the groupings of users, Web pages they are interested in and the tags they personally choose for these pages. Web page recommender system takes cluster numbers of users, resources as an input. Before recommendation, system generates candidate sets of Web pages for recommendation that are specific for each cluster. Candidate sets are generated by selecting most frequently tagged Web pages in the Web page clusters. The number of times a Web page is tagged by a user for each Web site is recorded in a dataset for this operation. For every Web page cluster, Web pages in it are ordered by the number of times they are tagged by users in general. By other words, Web pages are ordered by the count of people who decided to tag them. After ordering, the most frequently tagged ones are listed on the top of the list. The proposed model is capable of being configured to generate different number of recommendations for the users. Candidate sets are generated according to the number of recommendations. Each candidate set is filled with the Web pages, which are taken from the top of the ordered list with the preferred count of recommendations.

In the recommendation step, system looks for the user profile in the training set. User profile is determined simply by the cluster no of the user. After tripartite clustering each user is matched with a specific cluster, which includes the similar users with same interests, the Web pages they are interested in and the keywords they assigned to Web pages. If the system has a record of the matching cluster of this user, it means that the user is recognized by the user. If user is recognized, the Web pages in the candidate set that corresponds to the users cluster are recommended.

The Cold Start Problem

The social resource sharing systems include a registration step before use in general. Delicious Web site also has a registration system for keeping track of users and recording their bookmarks with some basic personal information. By other means,

the resource sharing system can recognize users easily because it requires registration before use. But still, a new user, who have just registered system and have not bookmarked any Web site yet, cannot be categorized by the recommendation model. The resource sharing system is able to identify the user since it asks for some basic personal information in the registration step, but the recommendation model can't identify his/her since the model have not any information about the interconnections with the resources and tags, which it works on. Without the links between the user, resources and tags, the model can't cluster his/her in the pattern discovery step and can't generate recommendations in the recommendation step. This situation is called the cold start problem in the literature. Cold start problem in recommender systems is the problem of making recommendations for new users [28].

Several methods are proposed for overcoming this problem in all kind of systems as well as recommendations systems. It is important to overcome this problem for the system, since it improves the efficiency of the recommendation model. In this work, we implement a simple way to overcome this; we offer a common set of most frequently tagged Web pages to users. Another candidate set for new users, which include the most visited Web sites in the main dataset and recommends the pages in this set, is generated. This method satisfies user needs by suggesting popular sites to unknown users and also improves the efficiency of the system. A reasonably better way for overcoming this problem on our system is explained in the Future Works section.

4.3 Tag Recommendation Model

Output of the tripartite clustering model also extracts the tag groupings in the dataset. Similar to Web page recommendation, first candidate sets of tags for each tag cluster is generated. These sets consist of the most frequently assigned tags in their clusters. All of the tags in the main dataset are ordered by the number of times they are assigned and recorded in a file, which is taken as an input by the system. The same tags can be might be assigned by different users, or they can be assigned for different resources by the same people. They are ordered by the count of times they are assigned, not by the number of people assign them or not by the number of resources they are assigned to. This kind of listing is capable of knowing the popularity of tags among both users and tags. This could be changed if it is decided that the

recommendation model would be user centric or resource centric. Or, the choice of ordering can be changed when the recommendation is made by analysing the user profile or the resource profile.

In the recommendation step, the resource profile is checked first. Resource profile is the cluster the resource is in. When the enormous size of Internet is considered, it is unlikely for any system to have a record for every resource. It is impossible because many new Web sites are constructed and set online every day. But still, since humans have a common sense of quality in general, it is expected that some of the resources are liked and preferred by most of the users. Some of the resources become more popular and social resource systems have record of these popular ones. When the system finds a record of the cluster of the resource, the model recognizes it and can retrieve its matching candidate set. The tags in the candidate set are then suggested to the user.

When the resource is not recognized by the system, a different type of cold start problem occurs. To overcome this problem, user profile is checked in the system. By other words, if the tags related to this Web site can't be obtained by the system, it retrieves the tags related to this user. We first implemented the model in the way that it first suggests tags using the user profile, until he/she bookmarks a Web page. Since this is an online process in a real time system, this is a reasonable approach. Other approaches are also possible for this, like retrieving a union of the cluster of user and the cluster of resource. In the final implementation of the model, resource profiles are analysed first since resource profiles provide more specific information then user profiles.

If the user is also not recognized by the system, a common set of tags from the main dataset is suggested. Most popular tags are recommended to the user.

5. EXPERIMENTAL STUDY

5.1 Information About Dataset

Original dataset is gathered from rss feed of Del.ici.ous Website and it was already ordered by tagging time. We divided the dataset into training and test sets using artificial time sessions in order to generate recommendations for user profiles. Original dataset contains 213,628 posts created by users. After applying pre-processing steps, the resulting dataset is divided into training and test datasets with % 70/% 30 ratios. Training set includes 12,998 rows, representing transitions between 1054 unique users, 900 unique resources and 668 unique tags. Each row corresponds to a tagging of a user in the following format, [user url tag] Test set, in the same format, contains 5,571 rows, representing transitions between 552 unique users, 629 unique resources and 579 unique tags.

5.2 Experimental Results

After data preparation and cleaning steps, the training set is used as an input for the Tripartite Clustering Model in the pattern discovery and analysis step of the recommendation process. Training set is also used in the recommendation step for generating frequency tables for the recommendation model. Experiments are conducted on the test set, by assuming that every line in the test set is a bookmarking operation on the real time system.

Since the goal of this thesis is to evaluate the impact of the tags on recommendation, results of the recommendation model are evaluated by comparing them with another recommendation model which is based on bipartite graph clustering. The other recommendation model is implemented in the same manner except the fact that it use resulting clusters of bipartite clustering. Still, the model has some differences since it can't use the tag information also in the recommendation step other than pattern discovery stage.

We use bipartite clustering of users and documents for Web page recommendation and bipartite clustering of documents and tags for tag recommendation in our recommendation model for comparison of experimental results. To extract the bipartite partitions in the same dataset, SRE algorithm is applied to the pairs. SRE algorithm [29] is a well known technique that implies recursive spectral graph clustering on bipartite structured data. Graph Analysis Toolbox [30] is used for bipartite graph partitioning operations.

5.2.1 Experimental Results for Web Page Recommendation

For evaluating the results, the similarity of the page, which is visited by the user, and the page, which we have recommended, is calculated. Users in a social bookmarking system are allowed to save and tag any Web site on the Internet. When the enormous size of the Internet is considered, it is unlikely to recommend the exact page that the user will visit, for any recommender system which is based on folksonomy data. This fact is also stated as a problem in other studies [20]. Semantic similarity of recommended page and visited page is a better measure to evaluate effectiveness of the system. Since tags of a Web page provide valuable information about semantics of it, the pages are compared using their tag vectors. Each resource is identified with a set of tags which includes the tags assigned to it by any user in the training set. For instance, the resources in the first column of Table 5.1 can be identified as the matching set of tags in column 2. Table 5.1 shows that tags of a Web page provides valuable information to identify it.

Table 5.1: Set of tags to identify resources.

Resource	Set of Tags
http://www.smugmug.com/	[photograph, blog,design, art, storage, image, web2.0, community, photosharing, hosting]
http://280slides.com/	[keynote, application, slideshow, apple, web2.0, web, productivity, tool, online, presentation, free, software, web, powerpoint, design, javascript, toread, slides, webdesign, inpiration, cloud]
http://www.revostock.com/	[video, audio, product, business, database, library, footage, music, media, resource, post, prodution, sell, project, stock, shopping, service, download]
http://smittenkitchen.com/	[cooking, recipe, baking, bagel, blog, photograpy, food]
http://thread.com/	[facebook, socialnetworking, dating, design, friends, social, startup, design, inspiration]
http://creately.com/	[tool, collaboration, modeling, mindmapping, chart, diagram, visualization, design, online, webdev, uml, webservice, flowchart, drawing, charts]
http://storybird.com/	[digitalstorytelling, interactive, writing, book, story, collaboration, onlinepublishing, children, english, web2.0, technology, literacy, publishing, online, free, elemantary, resource, art, visual, teacher]
http://dotsub.com/	[translation, subtitle, video, film, language, video, web2.0, tool, collaboration, community]
http://www.itsnicethat.com/	[design, blog, inspiration, photograppy, advertising, creative, illustration, culture, art, daily, magazine, graphic]
http://www.boostermp3.com	[mp3, music, download, search, free, audio, tool, searchengine, websearch, fun, online, recommendation, mplayer, search, streaming]

Tag vectors are used for computation of similarity metrics for evaluation. Cosine similarity and Jaccard similarity is calculated for analysing the similarity of resources. Jaccard similarity coefficient is a statistical measure of similarity between sets. For two sets, it is defined as the cardinality of their intersection divided by the cardinality of their union and it is represented as,

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad \textbf{(5.3)}$$

The other metric, used in the analysis, cosine similarity, is a measure of similarity between two vectors by measuring the cosine of the angle between them [31]. Given two vectors of tags, A and B, the cosine similarity is represented as follows,

$$similarity = \frac{A \times B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad \textbf{(5.2)}$$

A and B is the term frequency vectors of the tags. Cosine similarity is evaluating the similarity of documents using tag vectors, so it considers frequency of the tags while computing. In this context, for folksonomy, frequencies of the tags provide valuable information for defining a resource. For instance, a popular tag like 'design' is more definitive for a resource than a more personalized keyword like 'mystuff'.

In order to test the Web recommender model, a Web page is suggested for every bookmarking operation in the test set. When a user decides to tag a Web page, a specific number of recommendations are generated to the user. Each suggested Web page is compared with the visited Web page and, the Web page with the maximum similarity is selected as chosen by the user in the real time system. At the end of the evaluation, the mean value of the similarities for each Web site is calculated.

The SRE algorithm doesn't take the number of clusters as input but it allows configuring recursion capability and cut-off parameters. Since it decides the number of clusters in the dataset itself, for comparison we computed the Tripartite Clustering Model for same number of clusters.

Experiments are repeated for 10 times and mean±SD values are presented for tripartite since Tripartite Clustering Model starts with random initialization. Top 5 frequently used resources of the cluster, are recommended.

Table 5.2: Mean Similarity of recommended and visited pages.

Number of Clusters	Tripartite Clustering		Bipartite Clustering	
	Cosine Similarity	Jaccard Similarity	Cosine Similarity	Jaccard Similarity
12	0.223±0.01	0.358±0.01	0.209	0.358
13	0.226±0.01	0.363±0.01	0.212	0.360
14	0.226±0.01	0.359±0.02	0.213	0.360
19	0.226±0.01	0.359±0.01	0.225	0.372
22	0.228±0.01	0.354±0.01	0.228	0.375

Results of the experiments indicate that, for Web recommendation, usage of resulting clusters of tripartite clustering outperforms clusters of bipartite clustering.

Figure 5.1 presents tag precision values in %. Experiments are conducted for different number of clusters between 10-20. Since they resulted with close outcomes, we present the results with 13 clusters here. It can be seen that, the count of urls, which are related with the recommended page with only one tag, outperforms on bipartite clustering approach. But when overlap ratio of tags increases, tripartite clustering approach outperforms bipartite clustering approach. These results show that, recommendation with tripartite clusters, can suggest more relevant pages for the user. This figure shows the effect of involving the tags while clustering. Tags give the information about the purpose of the user while visiting the page and this helps to recommend more relevant results for his/her intention.

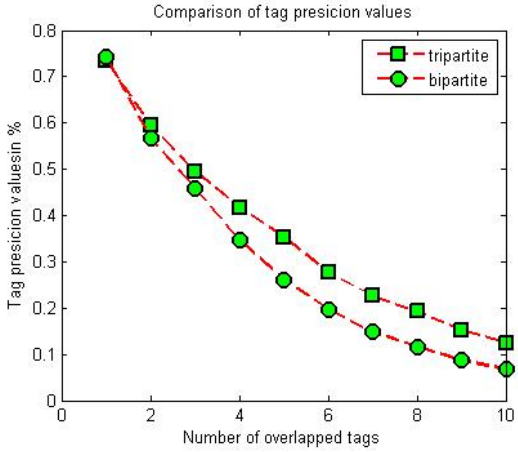

Figure 5.1 : Comparison of tag precision values in %.

5.2.2 Experimental Results for Tag Recommendation

Tripartite clustering also clusters tags using the resources associated to them and users who are interested in the topic. Resulting tag clusters can be used for recommending tag names to the user for the next Web page, using his/her interest. Top 5 frequently used tags of the cluster are recommended.

For comparison bipartite graph clustering is applied to resources and tags. Table 5.3 presents the results of the tag recommendation. As can be seen from the table, bipartite outperforms tripartite for recommending tag values.

Table 5.3: Mean similarity between recommended tags and assigned tags.

Number of clusters	Tripartite	Biparite
13	0.335±0.02	0.355
14	0.335±0.02	0.349
17	0.338±0.01	0.351
20	0.326±0.02	0.351
23	0.319±0.02	0.350

Experiments are conducted using WordNET Similarity [32] for calculation of semantic similarity of the tags. We were expecting that semantic similarity would be a better metric for pointing out the efficiency of tripartite over bipartite clustering but uncontrolled vocabulary of the tags restricts this kind of evaluation since corpus of WordNET can't cover this kind of vocabulary. We implement stemming to overcome this but we had troubles on finding the stemmed word in the corpus. So both the stemmed and one of the original versions of tags are compared with the recommended ones. Still there are many tags that are considered as dissimilar to the recommended one by WordNet. Table 5.4 presents some of these examples.

Table 5.4: Example of tags that are considered as semantically unrelated.

Assigned Tag	Recommended Tags
Regex	css jquery javascript Webdev develop wordpress
Php	css jquery javascript Webdev develop wordpress
Analytics	twitter internet socialmedia social market media
Podcast	research new fun humor culture movie
Photoshop	art photography illustrator graphicdesign inspiration architecture
Filesharing	software collaboration product storage calendar download

The reason why we compared the tripartite partitions of the dataset with the bipartite partitions was to evaluate the impact of the third attribute in clustering for recommendation. The results of Web page recommendation prove that it is reasonable to cluster the Web pages by taking into account the users' topic of interest on them. It is also expected that the third attribute for tag recommendation, the user information, would help the system to recommend tags that are more likely to reflect the users interest on the page. In contrast, our recommendation system is less successful than expected on tag recommendation. This may be the result of the existing tag recommendation system on Del.ici.ous Web site. [2] When the user wants to save a new bookmark, a form is presented to the user, which allows him to

add tags and notes related to the bookmark. Tag field also includes a recommendation line below, which suggest tags to the user. These recommendations are generated from an existing recommender system and since the dataset is crawled from a real-time system with an existing recommender, users in this set are more likely to choose one of the recommended tags rather than assigning a word from their own vocabulary. This would reduce the effect of user interest and corrupts the tripartite structure of folksonomies.

6. CONCLUSION

The major purpose of this research is to develop a recommendation system with better accuracy results for internet data by using folksonomies. For this, we implement a model that fits tripartite structure of folksonomies and extracts valuable information for generating recommendations. We compare our recommendation results with the results of a bipartite clustering model for both Web pages and tags. Results show that this model generates better accuracy results for Web page recommendation while extracting more useful information simultaneously which enables us to generate different types of recommendations.

Future work would be to examine the performance of the recommendation system on a larger dataset with different evaluation metrics. Also our recommendation system would be improved for cold start problem of new users. Several methods are proposed for overcoming this problem in all kind of systems as well as recommendations systems. It is important to overcome this problem for the system, since it improves the efficiency of the recommendation model. In this work, we implement a simple way to overcome this; we offer a common set of most frequently tagged Web pages to users. Another method to overcome this problem in a bookmarking site, which uses the recommendation model proposed in this thesis, would be to ask the user to choose at least one Web page and/or tag, which he/she might be interested in, on the registration step. The personal information gathered in the registration step might be used in many different ways to know more about the user but for this recommendation model, what it needs is some links between the users, resources and tags. The system needs to place the user to a cluster to identify her. At least one link between the user and the other node, resource or tag, would be enough for the user to be clustered. If the user only chooses a tag, the tag can be used to place the user into a cluster. This can be implemented in real systems by asking the users their general area of interests by providing a common set of tags for them to choose. This method is a better way for overcoming the cold start problem since it allows the recommendation model to generate more specific recommendations than suggesting

a common set of Web pages. The recommendation model would be reasonably more efficient when improved with this method.

The recommendation model can also be improved by suggesting union of the candidate sets of user and resource clusters for tag recommendation. In the current configuration of the system, candidate set of the resource cluster is first suggested. If resource cluster is not recognized, cluster of the user is considered. And if they are both not recognized, a common set of popular tags are recommended. The tag recommendations might result with better accuracy results if their candidate sets are generated by union of clusters rather than being user centric and resource centric. The tripartite clustering model already clusters the nodes by considering both users and resources and it also can configure the importance of user or resource nodes by changing a parameter in the formulation. Making this configuration in recommendation step, other than pattern discovery step, would improve the accuracy results of the system. Also, allowing the user to select his/her preference would enhance the model in a real time bookmarking system.

REFERENCES

[1] **Mathes, A.**, 2004. Folksonomies-Cooperative Classification and Communication Through Shared Metadata.

[2] <*http://delicious.com/* >, accessed at 15.08.2011.

[3] **Quintarelli, E.**, 2005. Folksonomies: power to the people.

[4] **Ciro Cattuto, Christoph Schmitz, Andrea Baldassarri, Vito D. P. Servedio, Vittorio Loreto, Andreas Hotho, Miranda Grahl, and Gerd Stumme**, 2007. Network properties of folksonomies. AI Commun. 20, 4 (December 2007), 245-262.

[5] **Melville, P. and Sindhwani, V.**, 2010. Recommender Systems., in Claude Sammut & Geoffrey I. Webb, ed., Encyclopedia of Machine Learning, Springer, 829-838.

[6] **Van Metern, R. and van Someren, M.**, 2002. Using Content-Based Filtering for Recommendation, Technical report, Foundation for Research and Technology - Hellas.

[7] **Ögüdücü, S. G.**, 2010. Web Page Recommendation Models: Theory and Algorithms, Morgan & Claypool Publishers .

[8] **Musto, C., Narducci, F., de Gemmis, M., Lops, P. and Semeraro, G.**, 2009. STaR: a Social Tag Recommender System, in Proc of the ECML PKDD Discovery Challenge 2009 (DC09), CEUR Workshop, Bled, Slovenia, 215-227.

[9] **Dattolo, A., Ferrara, F. and Tasso, C.**, 2010. The role of tags for recommendation: a survey, in Proc. of the 3rd International Conference on Human System Interaction - HSI'2010 , IEEE press, Rzeszow, Poland, 548-555.

[10] **Schmitz, C., Hotho, A., Jäschke, R. and Stumme, G.**, 2006. Mining Association Rules in Folksonomies, in Data Science and Classification, in Proc. of the 10th IFCS Conf., Springer, Berlin, Heidelberg , 261-270 .

[11] **Satoshi Niwa, Takuo Doi, and Shinichi Honiden**, 2006. Web Page Recommender System based on Folksonomy Mining for ITNG '06 Submissions, in Proc of the Third International Conference on Information Technology: New Generations (ITNG '06). IEEE Computer Society, Washington, DC, USA, 388-393.

[12] **Andriy Shepitsen, Jonathan Gemmell, Bamshad Mobasher, and Robin Burke**, 2008. Personalized recommendation in social tagging systems using hierarchical clustering, in Proc of the 2008 ACM conference on Recommender systems (RecSys '08). ACM, New York, NY, USA, 259-266.

[13] **Ching-man Yeung, C., Gibbins, N. and Shadbolt, N.**, 2007. Mutual Contextualization in Tripartite Graphs of Folksonomies, The Semantic Web, Springer, Berlin/Heidelberg, 966-970 .

[14] **Hwang, S.-H.**, 2007. A Triadic Approach of Hierarchical Classes Analysis on Folksonomy Mining.

[15] **Robert Jaschke, Andreas Hotho, Christoph Schmitz, Bernhard Ganter, and Gerd Stumme**, 2006. TRIAS--An Algorithm for Mining Iceberg Tri-Lattices, in Proc of the Sixth International Conference on Data Mining (ICDM '06). IEEE Computer Society, Washington, DC, USA, 907-911.

[16] **Dominik Benz, Andreas Hotho, Robert Jäschke, Beate Krause, Folke Mitzlaff, Christoph Schmitz, and Gerd Stumme**, 2010. The social bookmark and publication management system bibsonomy. The VLDB Journal 19, 6 (December 2010), 849-875.

[17] **Ching-man Au Yeung, Nicholas Gibbins, and Nigel Shadbolt**, 2007. Tag Meaning Disambiguation through Analysis of Tripartite Structure of Folksonomies, in Proc of the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Workshops (WI-IATW '07). IEEE Computer Society, Washington, DC, USA, 3-6.

[18] **Shilad Sen, Jesse Vig, and John Riedl.**, 2009. Tagommenders: connecting users to items through tags, in Proc of the 18th international conference on World wide Web (WWW '09). ACM, New York, NY, USA, 671-680.

[19] **Jäschke, R., Marinho, L., Hotho, A., Schmidt-Thieme, L. and Stumme, G.** 2007. Tag Recommendations in Folksonomies.

[20] **Yang Song, Ziming Zhuang, Huajing Li, Qiankun Zhao, Jia Li, Wang-Chien Lee, and C. Lee Giles.**, 2008. Real-time automatic tag recommendation, in Proc of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '08). ACM, New York, NY, USA, 515-522.

[21] **Lu Caimei, Hu Tony, Park Jung-ran**, 2010. Clustering tagged Web based on the tripartite network of folksonomy.

[22] **Peter Mika**, 2007. Ontologies are us: A unified model of social networks and semantics. Web Semant. 5, 1 (March 2007), 5-15.

[23] **Caimei Lu, Xin Chen, and E. K. Park**, 2009. Exploit the tripartite network of social tagging for Web clustering, in Proc of the 18th ACM conference on Information and knowledge management (CIKM '09). ACM, New York, NY, USA, 1545-1548.

[24] <*http://arvindn.livejournal.com/115182.html* >, accessed at 15.08.2011.

[25] <*http://www.delicious.com/help/json* >, accessed at 15.08.2011.

[26]<*http://www.datawrangling.com/some-datasets-available-on-the-Web*>, accessed at 15.08.2011.

[27] <*http://tartarus.org/~martin/PorterStemmer/]*>,accessed at 15.08.2011.

[28] **Schein, A. I., Popescul, A., Ungar, L. H. and Pennock, D. M.**, 2002. Methods and metrics for cold-start recommendations, in Proc of the 25th annual international ACM SIGIR conference on Research and development in information retrieval , ACM Press, New York, NY, USA , 253-260.

[29] **Hongyuan Zha, Xiaofeng He, Chris Ding, Horst Simon, and Ming Gu.**, 2001. Bipartite graph partitioning and data clustering. in Proc of the tenth international conference on Information and knowledge management (CIKM '01), New York, NY, USA, 25-32.

[30] **<***http://eslab.bu.edu/software/graphanalysis/* **>,** accessed at 15.08.2011.

[31] **Russell, M.**, 2011. Mining the social web, O'Reilly, Beijing, Farnham.

[32] **<***http://wordnet.princeton.edu/ >,* accessed at 15.08.2011.

CURRICULUM VITAE

Candidate's full name: Yonca ÜSTÜNBAŞ

Place and date of birth: Zonguldak/Turkey 04/11/1986

Permanent Address: Merdivenköy Mahallesi Bahariyeli Sokak Mengiroğlu Apartmanı Kat 4 Daire 11 Göztepe Kadıköy İstanbul

Universities and Colleges attended:
Computer Engineering, Dokuz Eylul University
2004-2008
Zonguldak Science High School
2001-2004

Publications:

- Y. Üstünbaş, Ş. Gündüz-Öğüdücü, "A Recommendation Model for Social Resource Sharing Systems Based on Tripartite Graph Clustering", International Symposium on Open Source Intelligence & Web Mining (OSINT-WM 2011) in Conjuction with (European ISI 2011), 2011.